Database Resources of the NCBI

Peter S. Cooper

email: cooper@ncbi.nlm.nih.gov
phone: 301-435-5951

updated 7/22/02

Online Resources

Course Resources:

Power Point Slides:	ftp://ftp.ncbi.nih.gov/pub/cooper/GradCourse2002/jax.ppt
Handout:	ftp://ftp.ncbi.nih.gov/pub/cooper/GradCourse2002/CooperGradCourse.pdf
Problem Set:	http://www.ncbi.nlm.nih.gov/Class/jax/ExpGenetics/

General:

NCBI Homepage:	http://www.ncbi.nlm.nih.gov
Site Map:	http://www.ncbi.nlm.nih.gov/Sitemap/index.html
About NCBI:	http://www.ncbi.nlm.nih.gov/About/
NCBI News:	http://www.ncbi.nlm.nih.gov/About/newsletter.html

GenBank:

GenBank Release Notes:	ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

Collaborating Nucleotide Databases:

EMBL:	http://www.ebi.ac.uk/
DDBJ:	http://www.ddbj.nig.ac.jp/

Entrez:

Entrez:	http://www.ncbi.nlm.nih.gov/Entrez/

BLAST:

BLAST Main Page:	http://www.ncbi.nlm.nih.gov/BLAST/
Stephen Altschul's Lectures on BLAST statistics:	http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

Frequently Asked Questions:	http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.html
BLAST program Guide:	http://www.ncbi.nlm.nih.gov/BLAST/producttable.html
BLAST tutorials:	http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html
BLAST Clients, Executables and Databases:	ftp://ftp.ncbi.nih.gov/blast/
NCBI Source Code:	ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/

NCBI Structures:

Structure Homepage:	http://www.ncbi.nlm.nih.gov/Structure/
Cn3D tutorial:	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dtut.html
CDD Search:	http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
CDart:	http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps

Genomic Resources:

Entrez Genomes:	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
Human Genome Resources:	http://www.ncbi.nlm.nih.gov/genome/guide/
Map Viewer	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch/
LocusLink:	http://www.ncbi.nlm.nih.gov/LocusLink/
UniGene:	http://www.ncbi.nlm.nih.gov/UniGene/
Human Genome Sequencing:	http://www.ncbi.nlm.nih.gov/genome/seq/
Human Genome BLAST:	http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html
Spidey:	http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/
ePCR (UniSTS):	http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi

Mouse Genome Resources:	http://www.ncbi.nlm.nih.gov/genome/guide/M_musculus.html
Trace archive Megablast:	http://www.ncbi.nlm.nih.gov/blast/mmtrace.html
Mouse Genome Sequencing:	http://www.ncbi.nlm.nih.gov/genome/guide/M_musculus.html
Mouse Genome BLAST:	http://www.ncbi.nlm.nih.gov/genome/seq/MmBlast.html

Other Databases:

SWISS-PROT	http://expasy.cbr.nrc.ca/sprot/
PIR	http://pir.georgetown.edu/pirwww/pirhome.shtml
PDB	http://www.rcsb.org/pdb/
PRF	http://www.prf.or.jp/en/

Email addresses:

General Help	info@ncbi.nlm.nih.gov
BLAST	blast-help@ncbi.nlm.nih.gov

Literature Reference List

General

Baxevanis, A. and Ouellette, B.F.F., eds. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* Second edition
New York: John Wiley & Sons. 2001. ISBN: 0-471-38391-0

Gibas, C. and Jambeck, P. *Developing Bioinformatics Computer Skills.*
Sebastopol: O'Reilly and Associates. 2001. ISBN:1-56592-664-1

Mount, D. W. *Bioinformatics: Sequence and Genome Analysis.* Cold Spring Harbor Laboratory Press. Cold Spring Harbor. New York. 2001. ISBN: 0-87969-608-7

Wheeler DL, et al. 2002. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res*. **30**(1):13-16. PMID: 11752242

GenBank Database

Benson DA, *et al*. 2002. Genbank. *Nucleic Acids Res.* **30**(1):17-20
PMID: 11752243

Ouellette BF, Boguski MS.1997. Database divisions and homology search files: a guide for the perplexed. *Genome Res*. **7**(10):952-5. PMID: 9331365

BLAST

Altschul SF, *et al*. 1990. Basic local alignment search tool. *J Mol Biol.* **215**(3):403-10. PMID: 2231712.

Altschul SF, *et al*. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17):3389-402. PMID: 9254694.

Altschul SF, *et al*. 1998. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci.* **23**(11):444-7. PMID: 9852764.

Schaffer AA, *et al*. 1999. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*.**15**(12):1000-11. PMID: 10745990.

Tatusova TA, *et al*. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.* **174**(2):247-50. PMID: 10339815.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol.* **7**(1-2):203-14.PMID: 10890397; UI: 20346451

Zhang Z, *et al*. 1998. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* **26**(17):3986-90. PMID: 9705509.

MMDB and Structures

Marchler-Bauer *et al*. 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30**(1): 281-283. PMID: 11752315.

Wang Y, et al. 2002. MMDB: Entrez's 3D structure databse. *Nucleic Acids Res*. **30**(1):249-252. PMID: 11752307.

Specialized Genomic Resources

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921

Jang W, *et al.* 1999. Making effective use of human genomic sequence data. *Trends Genet*. **15**(7):284-6. PMID: 10390628.

Pruitt KD Maglott DR,. 2001 RefSeq and LocusLink: NCBI gene centered resources. *Nucleic Acids Res.* **29**(1):137-140. PMID: 11125071.

Sherry, S.T., *et al*. 2000. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**(1):308-311. PMID: 11125122

Tatusov RL, *et al.* 2001 The COG database: new developments in the phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**(1):22-28. PMID: 11125040

Wolfsberg, T. *et al.* 2001. Guide to the draft human genome. *Nature* **409**, 824 - 826.

Sequence Database Notes

Types of Sequence Databases

When discussing sequence databases it’s useful to classify them by whether they contain **primary** or **derivative** data.

Primary data are reports of biological sequences from the investigators that determined them. Derivative data are produced from primary database records and add value by performing some kind of computational analysis or additional annotation, compilation or curation. We will first discuss primary sequences databases focusing on GenBank, the NCBI’s primary sequence database and then turn our attention to derivative databases again focusing on the NCBI’s products.

Primary Databases

In a primary sequence database record there are typically no additional data or interpretation added by the maintainers of the database. These collections are archives of sequence information in that they contain records that may be unmodified or updated from the time they were submitted. A second consequence of this archival nature is redundancy in the data; there may be many examples of the same target sequence in the database submitted by different researchers.

The earliest examples of primary sequence databases are protein sequence collections. The Protein Information Resource (PIR) at Georgetown University is a direct descendant of one of the early protein sequence collections. As DNA cloning and sequencing technology improved, the number of nucleotide sequences available quickly surpassed directly determined amino acid sequences. Today by far the largest primary sequence databases are nucleic acid sequence databases. In fact most protein sequences available today are based on conceptual translations of the coding regions on DNA sequences and are therefore derivative data.

The International DNA Sequence Database Collaboration

The most important primary DNA sequence databases today are the three members of the international DNA sequence collaboration: GenBank, the European Molecular Biology Laboratory (EMBL) database and the DNA Database of Japan (DDBJ). All three of these databases accept direct submissions of sequence data and are products of government-sponsored institutions in their respective countries. GenBank is produced and maintained by the National Center for Biotechnology Information at the National Institutes of Health in the United States, the EMBL database by the European Bioinformatics

Institute in the United Kingdom and DDBJ by the Center for Information Biology of the National Institute of Genetics in Japan. All of these entities maintain a presence on the World Wide Web that includes web browser-based access to data and tools for sequence analysis. The scope of these centers includes more than just their primary database product. All are active centers of computational biology and bioinformatics research and produce other products including many important derivative databases.

The discussion here will focus on GenBank. Most of most of what is said will apply to the other two databases as well.

GenBank

The GenBank database has its origins in the dim past when it was produced in bound volumes. As the number of sequences increased and computer technology advanced, the database was made available on CD-ROM and came with software for accessing the data. The CD-ROM version was discontinued in 1997 when the number of CDs required became prohibitive. Right now GenBank is available through the Internet on the NCBI ftp site (URL: ftp://ftp.ncbi.nlm.nih.gov/genbank/). On the NCBI ftp server the database is made available as full data releases every two months in the even numbered months of the year. Between releases daily updates are provided. For each release, important information including release statistics is in the Release Notes (URL: ftp://ftp.ncbi.nlm.nih.gov/genbank/gbrel.txt). The current release (June 2002) contains over 20 billion bases and more than 17 million sequences from over 140,000 species.

Data files and GenBank Division Codes

On the ftp site, the GenBank data are divided into series of sequence files. Originally each of these sequence files corresponded to one GenBank division. The GenBank divisions are identified by a three-letter division code, for example the BCT (bacterial) division or the PRI (primate) division. The days of one file per division are long past now; the EST division is split into more than 100 files. However all of the more than 17 million sequences in GenBank are still separated into a handful of divisions. A discussion of GenBank divisions is a helpful in describing the kinds of data in GenBank and is also useful in searching the data on the NCBI web site using the Entrez system. For the purposes of this discussion we'll recognize two kinds of GenBank divisions, traditional and special sequence divisions.

Traditional Divisions

The traditional GenBank divisions contain sequences that are determined to a high degree of accuracy (1 error in 10000) and often have extensive annotation about the biology or features of the sequence. At first glance these traditional divisions appear taxonomic in nature. Closer inspection shows the overriding purpose in establishing them initially was to create single files of reasonable size. Taxa were spilt or lumped to accomplish this. For example, the primate (PRI) and rodent (ROD) sequences were separated from the rest of the mammalian sequences (MAM) because there were a large number of primate and rodent sequences. On the other hand, the fungal and plant sequences were lumped into the PLN division because originally there were fewer of these. Likewise all of the invertebrate animal taxa were lumped into the INV division.

Sequence data can be submitted to these divisions through the web based BankIt form (URL: http://www.ncbi.nlm.nih.gov/BankIt/). More complex submissions can be prepared using the NCBI standalone tool Sequin (URL: http://www.ncbi.nlm.nih.gov/Sequin/index.html)

Special Divisions

With changes in DNA sequencing technology and strategies, a number of special GenBank divisions were established. These are not based on the source organism of the sequence but are based on the technique used to generate the sequence or the intended use of the sequence. Most of the special sequence divisions are products of high throughput projects and are usually submitted in large batches by single projects. The GSS, EST and STS sequences are also present in separate databases apart from GenBank: dbGSS, dbEST and dbSTS. The format of the records within these databases is quite different than that used in GenBank. Submissions to these divisions are handled through different procedures and different staff than traditional submissions.

First pass sequence divisions

The expressed sequence tag (EST) division and the genome survey sequence division (GSS) were established to hold first pass single read sequences that have little or no annotation. Because these data are single sequence reads, the amount of sequence in each record is limited, and is likely contain sequencing errors including frame shifts and base miscalls.

EST division

The EST division holds automatically generated partial cDNA sequences. These sequences are derived from arrayed cDNA libraries. For each clone in the library, only a single read is obtained from each end of the insert using the standard sequencing primers. Thus there can be two sequences in GenBank for each clone. In the case of directionally cloned libraries these will be the 5’ end of the cDNA and the 3’ end. Robots are often used to automate the process of sequencing these clones. Large numbers of clones can be partially sequenced very rapidly using this strategy. Good sets of EST data are available for a number of organisms. In fact, the EST division is the largest division of GenBank

(168 sequence files in release 130). Although largely unannotated and error prone, these data provide a rich source of information about the expressed sequences in a particular cell type tissue or ultimately the organism. The EST data are an important resource gene discovery and gene expression data. At the NCBI, the EST data have been used to generate a derivative database; UniGene that attempts to organize these data into gene based clusters.

GSS division

The GSS division contains data that are the genomic equivalent of the EST data; that is first-pass single-reads of genomic clones. The bulk of the data in the GSS division are derived from bacterial artificial chromosome (BAC) libraries. BACs are the large insert genomic clones that are used in complex genome projects like the human genome project. This is explained in more detail below when we describe the HTG division. Sequencing centers will produce preliminary reads for these clones sometimes as a prelude to producing more complete sequences. These surveys go in the GSS division. Another related category of sequence comes from the extension of sequencing primers onto the insert of the clone. These so called BAC end sequences are used to identify overlapping clones and creating tiling paths for assembling large genomic contigs. The GSS division also contains whole genome shotgun sequencing reads for some organisms, most notably certain protozoan parasites. These GSS sequences are important resources for genomic sequence for these organisms even in this unassembled form.

The High Throughput Genome Sequence Division

Many of the large-scale genome-sequencing projects rely on a strategy that has been called hierarchical shotgun sequencing. Genomic libraries are made in large insert BAC vectors. The clones from these libraries are arrayed and then subcloned into plasmid vectors. The resulting mini libraries are then randomly sequenced until enough sequence is obtained at high accuracy to assemble this shotgun sequence to generate the insert sequence of the clone. Even the early stages of this assembly process are useful. So that investigators can have access to these incomplete or draft sequences GenBank established the High Throughput Genome sequence (HTG) division. Within the draft or HTG sequences, GenBank recognizes different phases of completion. These phases are based on the degree of coverage and assembly of the sequence in the record. Phase 1 records have sufficient coverage to have several assembled regions (contigs). However the order and orientation of these is unknown, and there will still be gaps of unknown length in the sequence. Phase 2 records have progressed to the point that the order and orientation of the assemblies is known, but there are still gaps. As more sequence becomes available, the submitters update the records and the records will progress through the draft phases until the coverage and accuracy are sufficient for the sequence to move to phase 3. The record will also move from the HTG division into one of the traditional

GenBank divisions: A human sequence then would move to the PRI division. A fly sequence would move to the INV division. A zebrafish sequence would move to the VRT division. Even though the sequences within them are incomplete, the draft sequences are still useful. The NCBI assembly of the human genome depends on draft sequences; as of July 2001, about half of the human genome is still in the HTG division.

The Sequence Tagged Site (STS) Division

STS division records are mapping reagents. A sequence tagged site is essentially a recipe for amplifying a specific fragment of genomic DNA using the polymerase chain reaction (PCR). The records generally include a pair of primers and the sequence of genomic DNA they amplify. STS markers are designed based on the sequence of a known gene, an EST, an mRNA or genetic marker. These markers are commonly used in the technique of radiation hybrid (RH) mapping as a means of constructing a physical map of a genomic region. In RH mapping a cell line from the species of interest (human for example) is given a lethal does of radiation. One effect of this is to break the genome into fragments. The fragmented genomic DNA of the irradiated cells can be rescued by fusing the irradiated cells with those of a different species. The resulting hybrid cells variously retain and expel fragments of the foreign genome so that unique clones from the hybrid line have differing portions of the foreign genome. Genomic DNA isolated from these clones can then be tested by PCR with STS markers to the irradiated genome. The probability that markers occur together in the same hybrid clone is inversely related to the distance between them in the original genome. The pattern of amplification can thus be used to construct a physical map showing the relative positions of these markers. Since the genetic position of many of these markers is also known, radiation hybrid map positions can be integrated with genetic maps. Finally since STS markers are also sequenced based markers they can be mapped onto the assembled genomic sequence. The NCBI tool electronic PCR (ePCR) will search a sequence for the presence of markers from the STS division. This tool has been important in assembling the human genome sequence.

Other Special Divisions

The patent division (PAT) contains sequences provided by the US Patent and Trademark office. These sequences are not well annotated and not particularly useful even for patent claim investigation since GenBank cannot assure that this division includes all patents.

The contig division (CON) contains records that are instruction sets for assembling larger sequences. This division exists partly because GenBank has a 350 Kb limit for a single sequence. An example of a record in the CON division is the one containing instructions for assembling the *Escherichia coli* K12 genome from the < 350 Kb pieces in the BCT division.

The high throughput cDNA (HTC) division was recently created for draft cDNA records. Like the HTG division these sequences can be finished and then will move into the corresponding traditional division.

Derivative Sequence Databases

Protein only databases

Currently protein-only sequence databases like PIR, the Protein Information Resource (http://pir.georgetown.edu/)and SWISS-PROT (http://www.expasy.ch/sprot/)are essentially derivative database because the majority of protein sequences in them come from translations of nucleotide sequences. Both of these databases curate the protein sequences extensively and add additional annotations. These include comparing various examples of the protein sequences derived from primary sources. Both the SWISS-PROT data and PIR data are available at the NCBI.

NCBI Derivative Databases

UniGene

The UniGene database contains sequence similarity-based clusters of expressed sequences. Naturally, the richest sources of expressed sequences are the EST data. These data over represent the number of transcripts because highly expressed messages will be present many times within the data. The goal of the UniGene is reduce the EST data and to identify all transcripts for a particular organism. UniGene data sets are available for those organisms with substantial EST data. There are currently UniGene collections for eight animals (human, mouse, rat, cow, cow, clawed frog, fruit fly and mosquito) and five plants (Arabidopsis, rice, wheat, barley and maize). The UniGene data are a rich resource for gene discovery. Because EST libraries are tissue specific, UniGene data can be used as a resource for gene expression information. The NCBI Serial Analysis of Gene expression pages and the CGAP pages take advantage of this latter feature.

UniGene Build Procedure

Expressed sequences and coding regions from genes are clustered by sequence similarity. This is done in stages after removing mitochondrial, vector sequences and masking for repetitive elements. An important problem is cross clustering of sequences for different genes. Cross clustering can arise because the level of sequencing errors in EST sequences may approach the level of sequence divergence of members of the same gene family. One way to avoid this is to focus on the 3' untranslated regions first, since these are less well conserved than coding regions. After clustering using 3’ sequence alignments, clone based edges are added; this means adding 5' reads from clones whose 3' ends have already been clustered.

LocusLink and the RefSeq Project

Because of the tremendous growth in primary sequence data and the archival nature of these datasets, it can be difficult to identify the best sequence for a gene and in some cases even to find the sequence of interest because of confusing nomenclature problems. The LocusLink database attempts to solve some of these problems by collecting relevant links to sequences and other data in NCBI data as well as some outside databases. Each gene is assigned a stable unique identifier (Locus ID) and titles of entries are assigned based on the relevant genome nomenclature committee guidelines, the Human Genome Nomenclature Committee for human genes, the MGD Mouse Nomenclature Committee for mouse. These titles are also propagated to the UniGene database as well to standardize nomenclature for the clusters. LocusLink also tracks historical name for the genes. The current scope is fruit fly, human, mouse, rat, HIV1 and zebrafish.

RefSeq mRNAs and Proteins

A project that is intimately related to LocusLink is the generation of curated reference mRNA and protein sequences (RefSeqs) for the genes for the LocusLink entries. Collaborators supply information about which sequence is an appropriate representative for a gene. To generate a reference mRNA sequence, the best representative primary database sequence that has a full-length coding region is chosen. This record is used to create provisional RefSeq records. Essentially this provisional RefSeq is a copy of the of the database sequence but also includes several annotation enhancements: additional publications, aliases, LocusID number, MIM number, map information, and official gene symbol and name. These provisional RefSeqs are then subject to human review. This review process provides further enhancements to the RefSeq including extension of the using sequence data in other GenBank records, or the literature, correction sequencing errors, addition of additional publications and a summary of gene function. The final product represents a review article about the mRNA or protein. RefSeqs are available through LocusLink and are included in the Entrez and BLAST databases. RefSeq mRNA and protein sequences have distinctive accession numbers; NM_ followed by six digits for mRNA and NP_ followed by six digits for proteins.

Other NCBI Reference Sequences

There are several other kinds of reference sequences that are generated by projects at the NCBI.

Model Transcripts and Proteins

Closely related to the NM_ and NP_ RefSeqs are the model transcript and protein sequences. At this point these are generated only for the human genome by aligning the RefSeq mRNA to the corresponding genomic region. The genomic sequence that aligns is then used to create a model transcript and its corresponding translation. In many cases these sequences, do not match exactly the RefSeq mRNAs. This could be caused by assembly problems, sequencing error or true polymorphisms. These model sequences have distinctive accession numbers beginning with XM_ and XP_ and like the NM_ and NP_ RefSeqs are available through LocusLink, Entrez and through the BLAST databases.

Noncoding RNAs

Alignments of non-coding RNAs for GenBank are used to produce model transcripts. These are given accession numbers with the XR_ prefix.

Genome Assemblies

NCBI has created its own assembly of the human genome project data. The assembly consists of sets of contigs that are in turn built from assembling overlapping draft (HTG) and finished human sequence from GenBank. These large records are available through LocusLink, the Entrez system and can be searched as a BLAST database on the Human Genome BLAST page. Their distinctive accession numbers begin with NT_. There are also assembled finished sequences from the mouse genome project that are given NT_ accession numbers. However at this time there are no assemblies of mouse draft (HTG) sequence at the NCBI. The mouse genome project has produced a large amount of whole genome shotgun sequence (six-fold coverage). This has been assembled by the mouse genome sequencing consortium into contigs and super contigs. Assembled contigs are available as a part of the rodent division (ROD) of GenBank. Much larger supercontigs or scaffolds have been built by taking clone based information (plasmid paired end reads) into consideration. These scaffolds are available as NCBI RefSeqs with NW_ accession numbers.

Reference Genomic Records

The final type of RefSeq is a reference genomic record (NG_). These are created to serve as fixed sequence regions of genome. They are needed where the sequence and placement of a region is well known but difficult or impossible to assemble automatically. An example is the beta globin cluster on chromosome 11. Presently these are produced only for the human genome.

A summary of RefSeq accessions is given below.

RefSeq Accession	Type of record
NM_, NP_	Reference mRNA, translation
XM_, XP_	Model Transcript, translation
XR_	Noncoding mRNA transcript
NT_	contig
NW_	Whole Genome Shotgun Supercontig
NC_	Reference Chromosome
NG_	Reference Genomic